Understanding Tourist Behavior using Large-scale Mobile Sensing Approach: A case study of mobile phone users in Japan

Santi Phithakkitnukoon[†‡], Teerayut Horanont[‡], Apichon Witayangkurn[‡], Raktida Siri[§],
Yoshihide Sekimoto[‡], Ryosuke Shibasaki[‡]

[†]*Department of Computer Engineering, Faculty of Engineering, Chiang Mai University, Thailand*

[‡]*Department of Civil Engineering, School of Engineering, University of Tokyo, Japan*

[§]*School of Tourism Development, Maejo University, Thailand*

Abstract

This article describes a framework that capitalizes on the large-scale opportunistic mobile sensing approach for tourist behavior analysis. The article describes the use of massive mobile phone GPS location records to study tourist travel behavior, in particular, number of trips made, time spent at destinations, and mode of transportation used. Moreover, this study examined the relationship between personal mobility and tourist travel behavior and offered a number of interesting insights that are useful for tourism, such as tourist flows, top tourist destinations or origins, top destination types, top modes of transportation in terms of time spent and distance traveled, and how personal mobility information can be used to estimate the likelihood in tourist travel behavior, i.e., number of trips, time spent at destinations, and trip distance. Furthermore, the article describes an application developed based on the analysis in this study that allows the user to observe touristic, non-touristic, and commuting trips along with home and workplace locations as well as tourist flows, which can be useful for urban planners, transportation management, and tourism authorities.

Keywords:
Mobile sensing, mobility pattern, GPS location traces, tourist behavior.

1. Introduction

With recent advances in information and communications technology (ICT), the cities of today have become increasingly instrumented and interconnected. The activities and movements of urban dwellers are constantly measured and recorded by ubiquitous sensors embedded in urban systems (e.g., CCTV, building access systems, public Wi-Fi) as well as by personal electronic devices (e.g., mobile phones, laptops, tablets). A large number of individual digital traces is generated from which community and city-level behavioral signatures can be captured. Collectively, an image of urbanism of the real (physical) world can be reconstructed digitally. Consequently, an analysis of the characteristics of a city, its functionalities, and the behavior of its inhabitant can be performed, as reported in recent studies. For example, Phithakkitnukoon et al. [1] introduced a map that shows most probable activities in different areas of a city from which they found that people who work in the same industry (e.g., restaurant, retail, etc.) tend to have similar daily activity patterns, based on their analysis of connected cell tower locations (Call Detail Records

(CDRs)) of nearly one million mobile phone users. Longitudinal mobile phone locations can also help reveal interesting characteristics of human mobility, as indicated by Gonzalez et al. [2], who discovered that despite the diversity of our travel history, human beings follow simple reproducible patterns. Likewise, Song et al. [3] found that people are 93% predictable regarding where they go. Song et al. [4] later developed a model that reflects the tendency of people for commuting between fixed locations on a regular basis. Contributing to these efforts, Phithakkitnukoon et al. [5] showed that people's traveling patterns are influenced by the geography of their social ties.

Human mobility is one of the most important ecological and social challenges of the 21st century. People travel for different purposes, e.g., commuting and tourism. Commuting trips are typically repeated with unchanged routes; hence, such trips are relatively predictable. On the other hand, touristic trips are less predictable. As such, understanding tourist traveling behavior is important for urban planning, transport management, and tourism authorities. Today's pervasive technologies, such as mobile phones that have become an indispensable part of many people's lives, and as seen in recent research studies, with sensing capabilities that help turn the phone into become a personal sensor, collectively create a new sensing paradigm that incorporates humans as part of a sensing infrastructure. By taking advantage of the sensing capabilities of mobile phones, researchers are able to collect an unprecedented amount of fine-grained behavioral data from people. It offers a great advantage over conventional survey studies of tourist behavior.

Traditional tourist behavior studies tend to rely on surveys and questionnaires. For example, Alegre and Pou [6] used survey data gathered from 56,915 tourists over three years to study the length of stay at Balearic Islands in Spain. Gokovali et al. [7] analyzed three-week questionnaire data collected from 1,023 tourists to study the length of stay on vacations in Bodrum, Turkey. More recently, Wu et al. [8] studied the choice-making process of Japanese tourists based on survey data collected from 1,253 respondents in Japan.

In this work, we used large-scale mobile sensing approach to analyze tourist behavior. We analyzed GPS location traces of 130,861 mobile phone users in Japan collected for one year. The rest of this article will describe our approach in using GPS location records to detect tourists. From these records, we were able to perform tourist behavior analysis and demonstrate applications that can be useful for urban planners, transport management, and tourism authorities.

The main contributions of this work include the following:

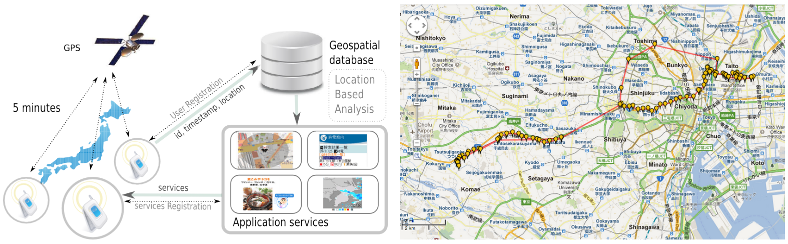
- a computational framework for identifying touristic trips from GPS location information (including algorithm for home and work location detection);
- a large-scale (country-level) analysis of tourist behavior, including touristic flows, time spent spent at destinations, choice of transportation mode, relationship between personal mobility and travel behavior, and similarity in travel behavior;
- a prototype application developed based on the analysis that allows the user to observe and analyze touristic trips and flows.

2. Identifying Tourists

2.1. Data

We capitalized on the opportunistic sensing paradigm that mobile phones can be personal sensors to use this personal communication device as a location tracker for our analysis. We used the GPS location records collected for a full calendar year (1 January 2012 - 31 December 2012) from 130,861 mobile phone users in Japan. The data was provided to us by one of the leading mobile phone operators in Japan, and collected from subscribers who registered for location-based services (and given consent for the use of their location information). The location information was sent through the network and used to perform specific analysis, from which certain services were then provided to the registered users, as shown in Fig. 1(a). To preserve user's privacy, the dataset was completely anonymized by the mobile phone operator before sending it to us. Each entry in the dataset included: unique user ID, position (latitude, longitude), timestamp, altitude, and approximate error (i.e., <100m, <200m, or <300m). To reduce battery consumption, an accelerometer was used to detect periods of relative stasis during which power-consuming GPS acquisition functions can be suspended. The sampling rate thus varied with the user's mobility but did not exceed once every five minutes. As an example, Fig. 1(b) shows location traces of a mobile phone user in our dataset.

Some of the subjects from our pool of subjects did not have GPS location traces for the entire year of our study for various reasons, such as the phone being turned off, not being subscribed to a service, or travel abroad; therefore, to ensure sufficient amount of data for our analysis, we selected the 130,861 subjects whose GPS locations were observed at least 350 days out of 365 days in 2012 (95%).



(a) Overview of data collection process

(b) An example of a mobile phone user's location traces

Figure 1: **Data collection process and an example data.**

2.2. *Home and workplace location detection*

In order to identify tourists from GPS trajectories, first we must detect the subjects' home and workplace locations to consider non-commuting trips only. It is from such non-commuting trips that touristic trips can be determined. To detect home and workplace locations, there are three steps in our approach, as depicted in Fig.' 2.

The first step was to identify *stop*, which was a collection of recorded GPS locations in close proximity, i.e., the location at which the user spends a considerable amount of time. A stop can be home, workplace, restaurant, market, etc. We recruited 15 subjects to carry a smartphone for one month with an application that allowed the subjects to identify stops that they made each day. With this ground truth information, we found that the spatial and temporal criteria [9] to identify *stop* most accurately were 196 meters and 14 minutes, as shown in our experimental results in Figs. 3(a) and 3(b). If $X_u = \{x_{t_1}, x_{t_2}, ..., x_{t_i}, ...\}$ denotes a set of GPS locations of user u where x_{t_i} is the location at time t_i, then our experimental results suggested that we group $x_{t_i}, x_{t_{i+1}}, x_{t_{i+2}}, ...x_{t_m}$ that are within 196m and $t_m - t_i \leq 14$min as a stop.

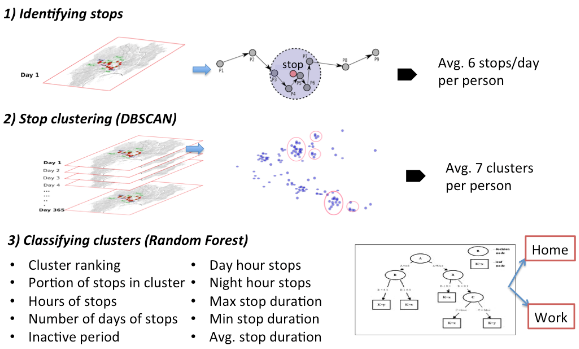
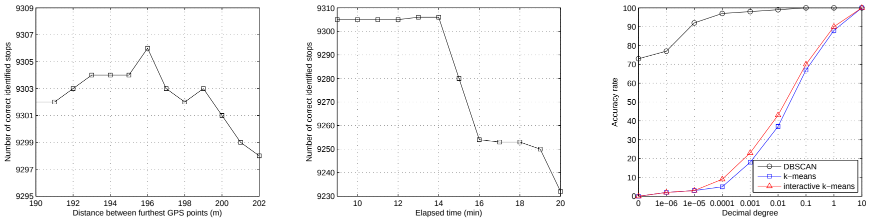


Figure 2: **Home and workplace detection method.**



(a) Stop detection accuracies for different distance thresholds

(b) Stop detection accuracies for different time thresholds

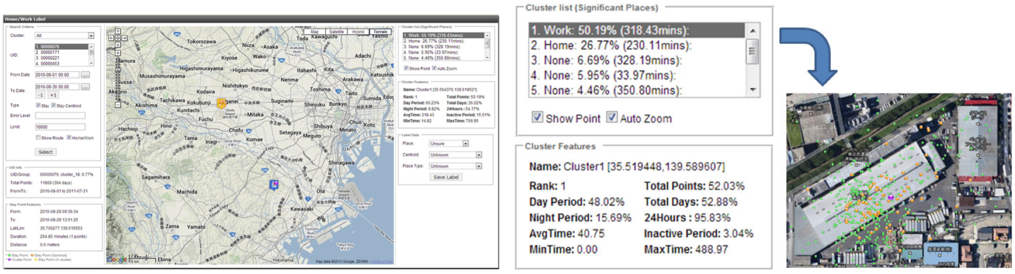
(c) Accuracy rates of different clustering techniques

Figure 3: **Experimental results for spatial and temporal criteria for stop.**

The second step was the spatial clustering of stops. The centroid of the cluster was to considered as a *significant place* (e.g., home, workplace, other). A number of clustering techniques were evaluated and we found that DBSCAN (Density-based spatial clustering of applications with noise) [10] had the best performance among other techniques. For validation, we developed a tool that allowed the tool user to label significant places after observing clusters of stops, as shown in Fig. 4. With this tool, we annotated our data with the home and workplace locations of 400 subjects, and used this as ground truth in our validation. Our experimental results (shown in Fig. 3(c)) indicate that DBSCAN outperformed other techniques (k-means, interactive k-means) in identifying the centroids of stop clusters that matched the locations of significant places. DBSCAN ($\epsilon = 30$ meters and MinPts $= 5$ points) achieved nearly 100% accuracy rate at 0.0001 decimal units ($\approx$11.1 meters), whereas k-means and interactive k-means can only achieve less than 10%. (Note: decimal degree of 1.0 is about 111.32 km.)

The last step was to classify significant places as home or workplace. The hand-labeling ground truth was used for this task and we found that Random Forest [11] had the best performance when compared with k-nearest neighbors and naïve Bayesian classifier (Table 1, 10-fold cross-validation was used) using the following 10 different features:

- *Cluster ranking*: top ranked clusters can be indicative of home and workplace locations.
- *Portion of stops in cluster*: to some extent, this suggests the importance of places because people tend to visit important places, such as home and work more frequently than others.



Cluster list (Significant Places)

1. Work: 50.19% (318.43mins):
2. Home: 26.77% (230.11mins):
3. None: 6.69% (328.19mins):
4. None: 5.95% (33.97mins):
5. None: 4.46% (350.80mins):

☑ Show Point ☑ Auto Zoom

Cluster Features

Name: Cluster1 [35.519448,139.589607]

Rank: 1	**Total Points:** 52.03%
Day Period: 48.02%	**Total Days:** 52.88%
Night Period: 15.69%	**24Hours :** 95.83%
AvgTime: 40.75	**Inactive Period:** 3.04%
MinTime: 0.00	**MaxTime:** 488.97

(a) Hand labeling tool

(b) Labeling example. Green dots are GPS locations, yellow dots are stops, and purple circle is cluster centroid.

Figure 4: **A snapshot of our developed hand labeling tool and labeling example.**

- *Hours of stops*: it is the portion of the hours of the day, where clustered stops appeared. For example, if stops were observed from 9am - 4pm (throughout the year), this feature would be 8/24.
- *Days of stops*: the number of days where clustered stops were observed.
- *Inactive hours*: for each subject, an inactive period was defined as the hours where a number of GPS locations is less than the average for at least three consecutive hours. Inactive-hours feature is a portion of clustered stops that fall into the inactive period.
- *Day-hour stops*: the portion of day hours (9am-6pm) that stops were observed.
- *Night-hour stops*: the portion of night hours (10pm-6am) that stops were observed.
- *Max stop duration*: maximum value of stop duration.
- *Min stop duration*: minimum value of stop duration.
- *Avg. stop duration*: average value of stop duration.

Table 1: **Performance comparison for classification of home and workplace**

Classifier	Home		Workplace		Other		Total	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Random Forest	87.32	87.57	89.94	77.28	94.19	98.58	90.48	87.18
k-NN	76.82	84.32	71.56	56.54	87.70	91.50	78.69	77.45
Naïve Bayes	71.93	84.91	78.75	71.36	96.35	94.50	82.34	83.59

To further validate our home location estimation, we compared our results against the census data and observed that the estimated population density based on our home location estimation was comparable ($R^2 = 0.966$) with the city population density information obtained from the 2006 Japanese Census [12], as shown in Fig. 5.

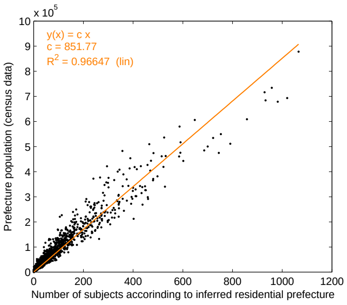


Figure 5: **Comparison between our inferred residential city population of mobile phone users and the actual city population obtained from the census data.**

2.3. Touristic trip detection

After obtaining a firm estimation of home and workplace locations, we were able to identify trips that were commuting (between home and workplace) as well as non-commuting. Suppose $S_u = s_{1,j}, s_{2,j}, ..., s_{i,j}, ...$ is a collection of stops of the user u where $s_{i,j}$ is the ith stop with spatial profile j; we defined a *trip* as a collection of stops that begins from home and ends at home. In other words, *trip* $= \{s_{m+1,home}, s_{m+2,j}, s_{m+3,j}, ..., s_{m+t,home}\}$ where t is the total number of stops in the trip. Therefore, a commuting trip is defined as a trip where at least one stop appears at a workplace ($s_{i,work}$), i.e., *commuting* $= \{s_{m+1,home}, s_{m+2,j}, ..., s_{m+i,work}, ..., s_{m+t,home}\}$. On the other hand, a non-commuting trip is defined as a trip where none of the stops appear at a workplace i.e., *non-commuting* $= \{s_{m+1,home}, s_{m+2,j}, ..., s_{m+i,j}, ..., s_{m+t,home}\}$.

From non-commuting trips, we identified touristic trips as a trip where at least one stop appears at a touristic destination i.e., *touristic trip* $= \{s_{m+1,home}, s_{m+2,j}, ..., s_{m+i,touristic}, ..., s_{m+t,home}\}$. For our analysis, we used the touristic destinations information provided by the Ministry of Land, Infrastructure and Transport of Japan (MLIT) [13]. The touristic destination information is composed of two sets of data: one contains location (latitude, longitude) of destinations, and the other contains polygons (i.e., shapefiles) of destinations. These two datasets do not overlap; in other words, the two datasets do not contain the same destination locations. To classify stops into touristic or non-touristic, we used both sets of data. We defined a touristic stop as a stop that is either within 200m from a touristic destination location or within the polygon area covered by a touristic destination. Hence, a spatial profile j can be identified as home, work, touristic, or non-touristic

3. Tourist Behavior

The detected touristic trips were then used in our analysis of tourist behavior. Based on our home location estimation, the number of subjects from 47 different prefectures of Japan is shown in Fig. 6, and the subjects' spatial distribution is shown in Fig. 7. The corresponding prefecture names are listed in Table 2. Tokyo has the highest number of subjects (15,586) followed by Kanagawa (8,588) and Saitama (7,329) as expected (because our subject distribution correlates well with the census population density (Fig.5)).

In our tourist behavior analysis, we were interested in *trip flows* – the number of touristic trips made to and from different prefectures in Japan, *time spent at destination*, *modes of transportation* used by the tourists, and correlations between *personal mobility and touristic travel* *behavior*. In great part the environmental and social effects of tourism are a product of the sheer volume of tourist trips [14]. Therefore, it is important to understand these characteristics of tourist flows in order to make better informed decisions as we move toward sustainable tourism.

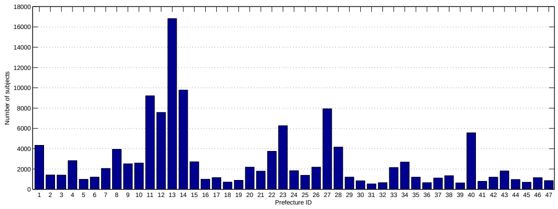


Figure 6: **Number of subjects from different prefectures (names are given in Table 2).**

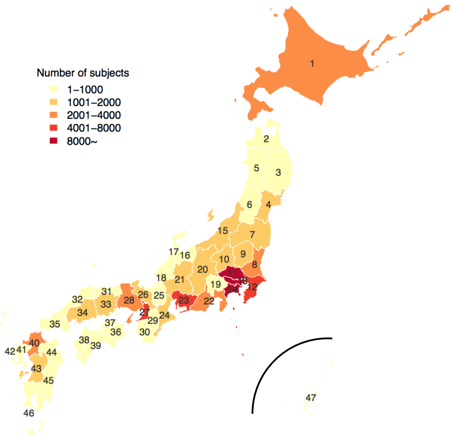


Figure 7: **Spatial distribution of number of subjects.**

3.1. Number of touristic trips

For each subject in our dataset, touristic trips were extracted using the algorithm described previously. The distribution of the number of trips made by the subjects is depicted in Fig. 8(a). There were a total of 1,987,858 trips made, which was equivalent to an average of 15.19 trips per person. Trips were classified into three categories: short (less than 5km), medium (5km-100km), and long (farther than 100km). The distribution of the number of trips in these three different ranges is shown in Fig. 8(b).

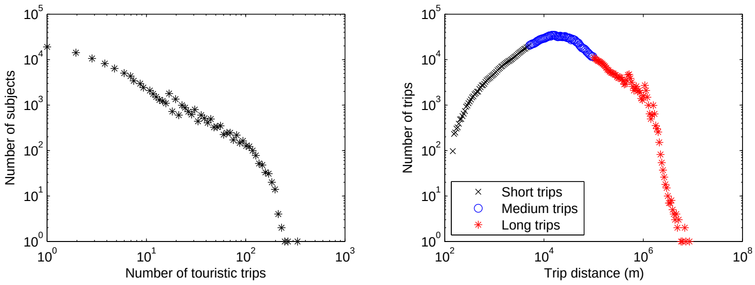
Based on the estimated home locations of the subjects, Fig. 9(a) shows the (sorted) outgoing touristic trips from different prefectures, where Tokyo has the highest outflow volume (495,904) followed by Kanagawa (161,422) and Fukuoka (109,962). The portion of trips made made corresponding to trip distance (short, medium, long) is shown in Fig. 9(b). Medium trips appeared to be the majority of the trips made in all prefectures. Likewise, by considering the visited prefectures as destinations, incoming flows are shown in Fig. 10(a) where Tokyo appeared as the top destination (1,754,902) followed by Kyoto (341,911) and Fukuoka (278,886). For the inflow volumes, the majority of the trips were a mixture of short and medium range trips, as shown in Fig. 10(b). One of the most important components considered in transportation management and engineering is the Origin-Destination matrix (O-D matrix), which describes trip distribution

Table 2: **Prefecture identifications and corresponding names**

Prefecture ID	Prefecture Name
1	Hokkaido
2	Aomori
3	Iwate
4	Miyagi
5	Akita
6	Yamagata
7	Fukushima
8	Ibaraki
9	Tochigi
10	Gunma
11	Saitama
12	Chiba
13	Tokyo
14	Kanagawa
15	Niigata
16	Toyama
17	Ishikawa
18	Fukui
19	Yamanashi
20	Nagano
21	Gifu
22	Shizuoka
23	Aichi
24	Mie
25	Shiga
26	Kyoto
27	Osaka
28	Hyogo
29	Nara
30	Wakayama
31	Tottori
32	Shimane
33	Okayama
34	Hiroshima
35	Yamaguchi
36	Tokushima
37	Kagawa
38	Ehime
39	Kochi
40	Fukuoka
41	Saga
42	Nagasaki
43	Kumamoto
44	Oita
45	Miyazaki
46	Kagoshima
47	Okinawa

and flow. In our case, O-D matrix can be derived from the touristic trip origin and destination information and it is shown in Fig. 11 from which we can observe the top origin and destination prefectures, as well as the trip distribution.

From the O-D matrix, it can be observed that the number of trips made to and from the same prefectures is relatively high, which is intuitive. A high number of trips can also be noticed between nearby prefectures within the same region, for example, the Northeast region (including Hokkaido and Tohoku region, prefecture ID 1-7), the Kanto region (prefecture ID 8-14), the Kansai region (prefecture ID 24-30), and the West region (including Okinawa and Kyushu region, prefecture ID 40-46). The O-D matrix also shows that prefecture population density seems to be indicative of outflow volume, but not inflow volume; for example, the number of trips made from highly populated regions, such as the Greater Tokyo region (Saitama, prefecture ID 11), Chiba (prefecture ID 12), Tokyo (prefecture ID 13), and Kanagawa (prefecture ID 14), is significantly higher than other prefectures, but the inflow volume is not as high as the outflow, particularly for Saitama and Chiba. On the other hand, Tokyo and Kanagawa have a much higher inflow volume because there are various touristic attractions in their prefectures. Tokyo is the capital city of Japan; Kanagawa's capital city is Yokohama, which has the largest Chinatown in Japan and one of the largest in the world, and Kamakura, another city in Kanagawa, is famous for Buddhist temples and Shinto shrines. Our results show that Kanagawa attracts a great number of tourists from Tokyo. Other examples are Osaka and Kyoto, prefectures in close location to each



(a) Distribution of number of trips made

(b) Distribution of trip distance of different ranges: short (less than 5km), medium (5km-100km), and long (greater than 100km)

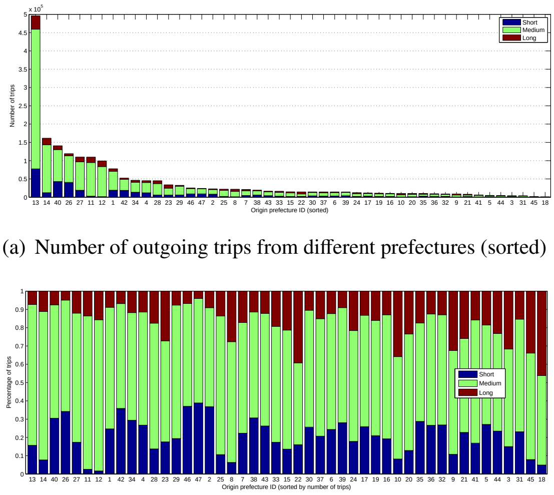
Figure 8: **Distribution of number of trips and trip distance.**

other that create flows in opposite directions; as a higher populated prefecture, Osaka (2,666,371 inhabitants) has a much higher outflow than the nearby prefecture of Kyoto (1,474,473 inhabitants). On the other hand, Kyoto attracts a much higher inflow than Osaka because of its main touristic attractions, which include various World Heritage sites. Inflows and outflows of Osaka and Kyoto are shown in Fig. 12. Another interesting observation from the O-D matrix is the flow between Yamaguchi (prefecture ID 35) and Fukuoka (prefecture ID 40), which are prefectures close to each other separated by the Kanmon Straits (a stretch of water that separates Japan's main islands). A trip by train from Hakata in Fukuoka to Yamaguchi takes approximately one hour; Yamaguchi has a beautiful town that flourished as "Kyoto in the West" during the medieval period, and it is one of the top touristic locations in the region. The O-D matrix shows large flows between the two prefectures, with a slightly larger flow from Fukuoka to Yamaguchi than in the opposite direction.

3.2. Time spent at destination

The amount of time tourists spend at their destination is another important aspect of tourist behavior. From the extracted touristic trips, we calculated the time spent at destination simply as the total amount starting from the time of arrival at the first destination until the departure time of the last destination of a trip. Tourists can visit more than one destination in one trip; consequently, we considered all destinations on a single trip as the overall destination on which the tourist spent time. The distribution of the time spent on a trip is shown in Fig. 13(a); the average was 543.40 min (≈9 hours) and the standard deviation was 764.39 min (≈12.5 hours). With different lengths of the trip, Fig. 13(b) shows the distribution of the likelihood of time spent at destination (normalized by the total number of trips) in short, medium, and long trips; from this figure it can be observed that overall, tourists are more likely to spend more than 546.83 min (≈9 hours) on longer trips.

To examine geography along with time spent by tourists, Fig. 14(a) shows the average time spent at destinations by tourists from different origin prefectures. It was observed that on average, tourists from Niigata spent most time at destinations on a trip compared to tourists from other prefectures (733.66 min ≈12 hours), followed by Nagasaki (696.47 min ≈11.5 hours) and Shimane (664.69 min ≈11 hours). Conversely, Fig. 14(b) shows the average time spent at different destination prefectures. Nagasaki and Hokkaido appeared to be the top prefectures that



(a) Number of outgoing trips from different prefectures (sorted)

(b) Percentage of short, medium, and long outgoing trips made from different prefectures (sorted by total number of trips)

Figure 9: **Trips made from different prefectures.**

tourists spent most time on their destinations (327.17 min and 326.31 min ≈5.5 hours) followed and Tokyo (312.82 min ≈5 hours).

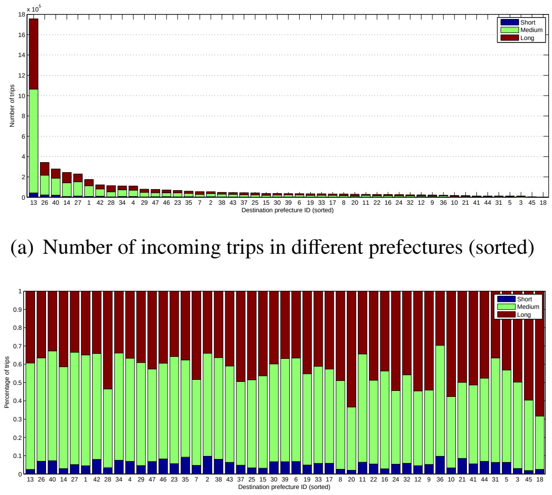
In case that there is a doubt in the differences from the results shown in Figs. 14(a) and (b), we would like to note that the average time spent at destinations by tourists from different prefectures shown in Fig. 14(a) is higher than the average time spent at different destination prefectures (Fig. 14(b)) because when considering tourists from different origin prefectures, tourists can visit more than one destination, which makes the total time spent on different destinations on their single trips higher than the time spent at particular destinations (in different prefectures), which is shown in Fig. 14(b).

Different types of destination can attract different tourists. Using the information provided by the Ministry of Land, Infrastructure and Transport of Japan [13], 26 touristic destination types were considered in our analysis. The list of these touristic destination types is listed in Table 3. All destinations detected in our data were labelled as one of these 26 touristic destination types. With this information, Fig. 15 shows the number of trips made to different types of destination. Shrine and temple, Building, and Annual event were the top types of destination that drew a large number of tourists. Figure 16 shows the average time spent at different types of destination; from this figure, it can be observed that Open field was the top destination type on which tourists spent most of their time (445.89 min ≈7 hours) followed by Botanical garden and aquarium (378.51 min ≈6 hours) and Museum (311.78 min ≈5 hours).

3.3. Mode of transportation

After exploring the number of trips to gain a better understanding of trip distribution and flow, as well as time spent at destinations by area and type, we extended our analysis to investigate another important aspect of traveling: the mode of transportation used by the tourist.

In our analysis of this study, we used the framework that we previously developed [15] for identifying modes of transportation used by mobile phone users based on their GPS locations.



(a) Number of incoming trips in different prefectures (sorted)

(b) Percentage of short, medium, and long incoming trips in different prefectures (sorted by total number of trips)

Figure 10: **Trips made from different prefectures.**

Our previously developed framework basically looked at the GPS traces along with detected stops, i.e., $\{s_{1,j}, x_t, x_{t+1}, x_{t+2}, ..., x_{t+m}, s_{2,j}, ...\}$ and defined a *segment* as a series of GPS locations between adjacent stops i.e., $\{x_t, x_{t+1}, x_{t+2}, ..., x_{t+m}\}$. These segments were then classified into walking and non-walking segments based on the rate of change in velocity and train line proximity. The walking segments were inferred as *walking* when used as a mode of transportation. The non-walking segments are then classified into two modes of transportation, *car* and *train* , based on Random Forest classification technique using the following features: segment distance, time elapsed in segment, speed (minimum, maximum, average, maximum acceleration, and rate of change in velocity), and portion of GPS points that fall into train and road networks. The framework was validated against the ground truth modes of transportation used by 100 subjects who were recruited for the study to carry a smartphone with an app that allowed them to input the transport modes they used daily over one month. Our framework was able to perform with an overall precision rate of 86.89% and a recall rate of 84.17%.

The modes of transportation considered in this study therefore were walking, car, and train. When tourists travel, they can employ a mixture of transportation modes in a single trip. Therefore, in our analysis, we considered the portion of *time* spent on different transport modes on a touristic trip, as well as the portion of *distance* traveled with different modes. The overall distribution of the time portion on different transport modes in a single trip is shown in Fig. 17(a). From this figure, we can observe that on average tourists spent slightly more time in a car (avg. = 48.57%) than walking (avg. = 45.21%) but much more time in a car and walking than on a train (avg. = 6.22%). For short trips, tourists (on average) tended to spend more time walking (avg. = 78.98%) than in a car (avg. = 14.89%) and on a train (avg. = 6.14%) as shown in Fig. 17(b). For medium trips (Fig. 17(c)), tourists were likely to spend more time in a car (avg. = 53.87%) than walking (avg. = 39.35%) and on a train (avg. = 6.78%). Likewise for long trips tourists: on average, tourists spent more time in a car (avg. = 64.04%) than walking (avg. = 33.11%) and on

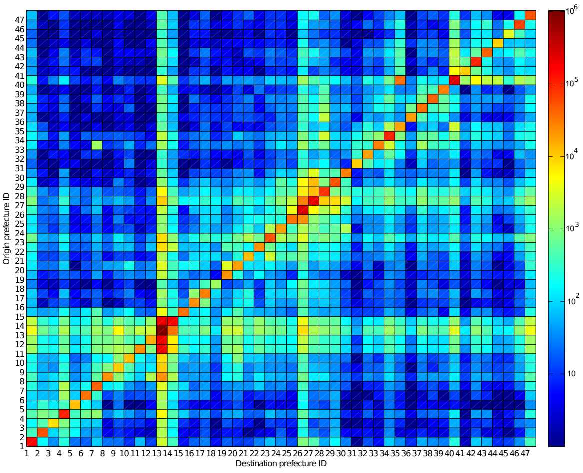
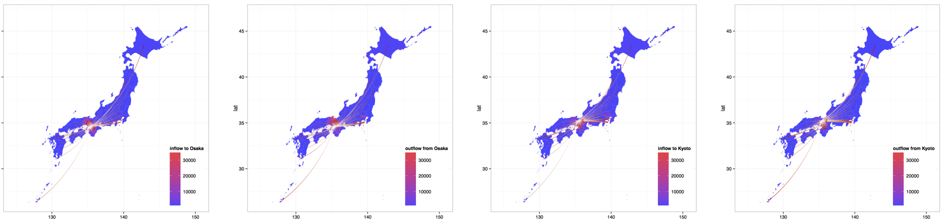


Figure 11: **Origin-Destination matrix (number of detected touristic trips made to/from different prefectures).**

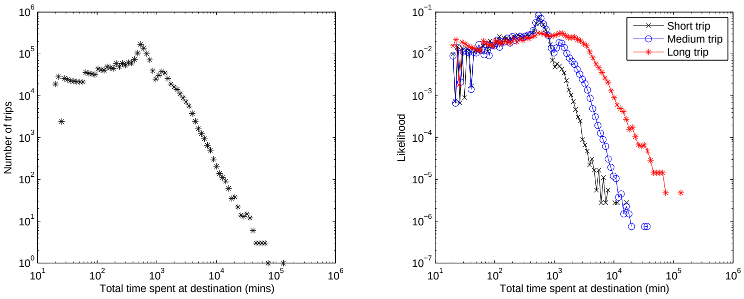


(a) Osaka inflow (b) Osaka outflow (c) Kyoto inflow (d) Kyoto outflow

Figure 12: **Touristic inflow and outflow of Osaka and Kyoto.**

a train (avg. = 2.85%) as shown in the distributions in Fig. 17(d). Values of mean and standard deviation of the results in Fig. 17 are listed in Table 4. Overall, as trip become longer, more time is spent in a car, but less walking; in comparison, time spent on a train remains the least among all transportation modes.

The distribution of the portion of distance traveled on a trip using different transport modes is shown in Fig. 18(a). We observed that on average the portion of distance traveled by a car (avg. = 71.51%) was higher than by walking (avg. = 2.52%) and by train (avg. = 6.97%). For short trips (Fig. 18(b)), tourists traveled by car (avg. = 72.46%) more than by walking (avg. 20.42%) and by train (avg. = 7.12%). For medium trips (Fig. 18(c)), it was observed that tourists traveled by car (avg. = 79.50%) for longer distances than by walking (avg. = 12.67%) and by train (avg. = 7.83%). Likewise for long trips, tourists tended to travel for longer distance in a car (95.33%) compared to walking (3.27%) and on a train (1.40%). Overall, on average, cars are used for traveling on a trip more frequently than trains and walking; as trips become longer, cars are used to travel longer distances, but distance traveled by walking decreases; in comparison, trains remain the least used mode of transportation among others. Means and standard deviations are listed in Table 4.



(a) Distribution of amount of time spent on a trip

(b) Likelihood of amount of time spent on a short, medium, and long trip

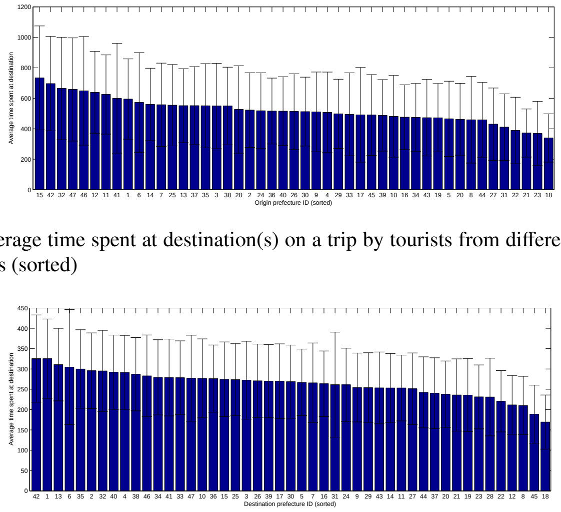
Figure 13: **Distribution and likelihood of amount of time spent on touristic trips.**

3.4. Personal mobility and tourist travel behavior

Can we predict or estimate tourist travel behavior, such as travel frequency, travel distance, time spent at destinations, and type of destination more likely to be visited, given the knowledge of tourists' personal mobility patterns? We wanted to answer this very question in our study.

We examined personal mobility using three matrices namely *mobility level*, *travel dispersion*, and *travel scope*. Mobility level of a person was defined as the total number of stops, which reflects activeness or how much the person travels. To measure how a person's mobility is spread out or dispersed, we used travel dispersion, which can be computed as $\sqrt{\sigma_{lat}^2 + \sigma_{long}^2}$ where σ_{lat} and σ_{long} are the standard deviation of points along latitude and longitude, respectively. When computing standard deviation i.e., $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$, the quantity $(x_i - \bar{x})$ i.e., distance from the mean was computed using the harvesting formula such that the unit of the measure was in kilometers. In other words, the travel dispersion can be interpreted as the magnitude of the resultant vector of the standard deviation of latitude and longitude. The last matrix used in quantifying personal mobility was travel scope, which was simply defined as the farthest destination that a person has even visited. In other words, the farthest destination was the farthest GPS location point from the person's home location. In this study, we only considered destinations in Japan; therefore, the farthest points considered in this study were restricted within Japan only.

We first examined the relationship between the mobility level and *travel behavior*, which was characterized by *the number of trips made*, *trip distance*, and *time spent at destination*. The results for the mobility level are shown in Fig. 19. Mobility level (ML) was considered for two types; touristic and non-touristic. The results show a strong relationship between touristic ML and the average number of trips with r^2=0.902 (α=0.247 for the linear regression of the form $y = \alpha x$), which is intuitive ; in other words, a higher touristic mobility level suggests that the person is likely to make higher number of trips. For the non-touristic mobility level, the results show strong inverse relationship with the number of trips for ML>200 (r^2=-0.745, α=-0.012) and for ML≤200 (r^2=-0.878, α=-0.348). When examining the relationship between touristic mobility level and trip distance, the results show a strong inverse relationship when ML≤200 (r^2=-0.953, α=-574.33) suggesting that the higher the touristic ML, the shorter the trip distance when ML is less than 200 stops. Touristic ML shows a strong relationship with time spent at destination (r^2=0.852, α=1.418); in other words, the higher touristic ML implies a longer time



(a) Average time spent at destination(s) on a trip by tourists from different prefectures (sorted)

(b) Average time spent at destination prefectures (sorted)

Figure 14: **Average time spent at destinations by tourists from and to different prefectures.**

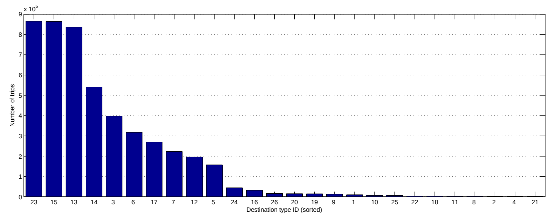


Figure 15: **Number of trips made to different types of destinations (sorted).**

spent at destination (on average). On the other hand, non-touristic ML was found to have a strong inverse relationship with time spent at destination when ML>200 (r^2=-0.735, α=-0.249) and when ML≤200 (r^2=-0.883, α=-3.126) – suggesting that the higher non-touristic ML implies less time spent at destination. For destination types, Figs. 19(g) and (h) show the top destination types and their corresponding percentages among other types across different levels of touristic and non-touristic ML. For touristic ML, we observed that Building was the top destination type followed by Annual Event, and Shrine and Temple across all levels of ML. For non-touristic ML, Shrine and Temple was the top destination type followed by Annual Event and Building. Shrine and Temple became a clear dominant type when ML was is approximately 300 stops.

For travel dispersion (TD), as shown in Fig. 20, we found a strong inverse relationship between touristic TD and and the number of trips (r^2=-0.862, α=-0.077), which suggests that the higher the touristic TD (i.e., touristic destinations are more spread out), the lower the number of trips. We also observed a strong relationship between touristic TD and trip distance (r^2=0.957, α=6457.6) when touristic TD≤80 km. When touristic TD≤22km, Shrine and Temple was found to be the top destination type followed by Annual Event but when TD>22km Annual Event was observed to be the top type followed by Building.

Table 3: **Touristic destination types**

Destination type ID	Type
1	Botanical garden and aquarium
2	Zoo
3	Museum
4	Open field
5	Historic site
6	National landscape
7	Castle
8	Mountain
9	Cave
10	Cape
11	Canyon
12	Island
13	Annual event
14	Garden and park
15	Building
16	Vegetation
17	Historic landscape
18	River
19	Coast
20	Lake
21	Wetland
22	Waterfall
23	Shrine and temple
24	Shrine garden and park
25	Natural phenomena observatory
26	Plateau

For travel scope (TS), as shown in Fig. 21, we observed a strong relationship between non-touristic TS and trip distance (r^2=0.866, α=0.391), which interestingly suggests that larger non-touristic TS implies a longer trip distance. In other words, the result tells us that people who visit places located farther away for non-touristic purposes are more likely to make longer touristic trips. For destination types, Annual Event appeared to be the top type overall followed by Building. Shrine and Temple was observed as top destination when TS<500km. For non-touristic TS, Building was the top destination type followed by Shrine and Temple, and Annual Event.

All values of r^2 and linear regression constant α of the results shown in Figs. 19, 20, and 21 are listed in Table 6.

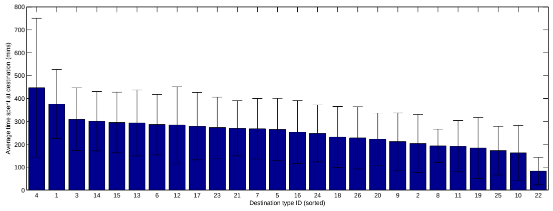
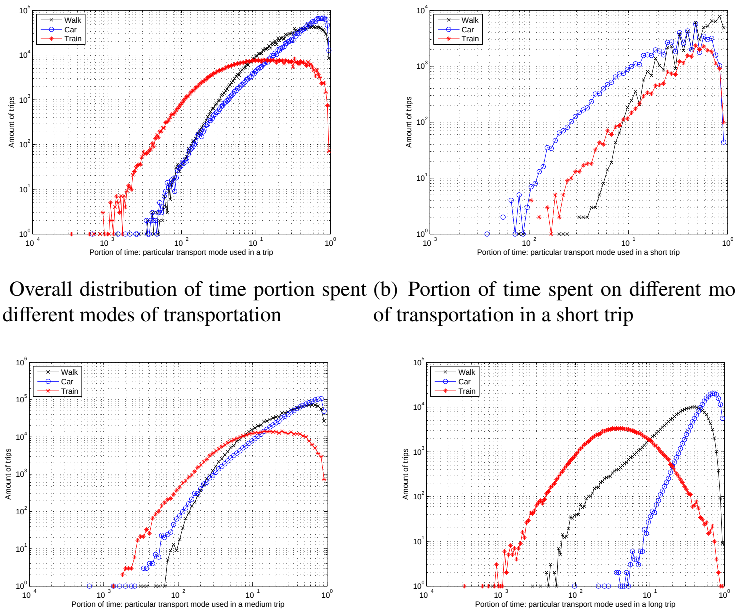


Figure 16: **Average time spent at different types of destinations (sorted), with standard deviation bars.**

Table 4: **Means and standard deviations of percentage of time on a trip spent on different transport modes (Fig. 17)**

Trip type	Mode of transportation					
	Walking		Car		Train	
	Mean	Std.	Mean	Std.	Mean	Std.
Short	78.98	35.50	14.89	30.36	6.14	20.56
Medium	39.35	27.91	53.87	29.10	6.78	15.57
Long	33.11	17.66	64.04	17.86	2.85	4.90
Overall	45.21	32.19	48.57	32.17	6.22	15.76



(a) Overall distribution of time portion spent on different modes of transportation

(b) Portion of time spent on different modes of transportation in a short trip

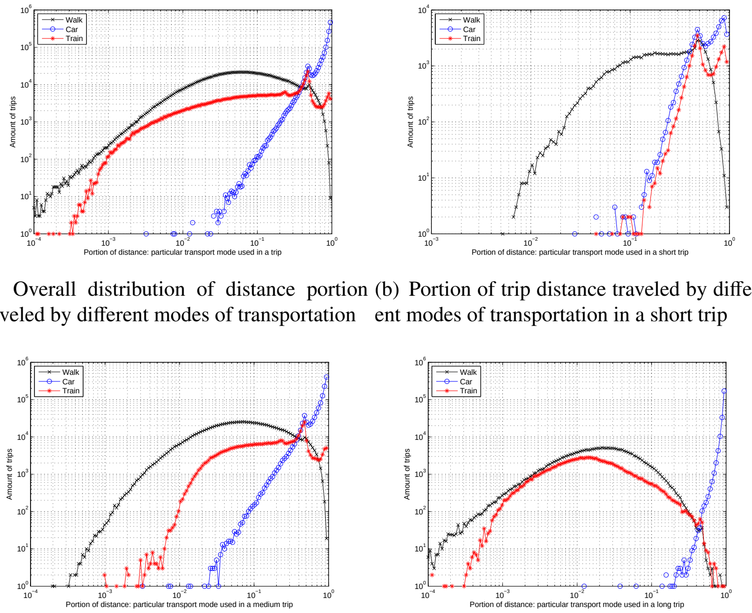
(c) Portion of time spent on different modes of transportation in a medium trip

(d) Portion of time spent on different modes of transportation in a long trip

Figure 17: **Distributions of portion of time spent on train, car, and walking in a trip.**

Table 5: **Means and standard deviations of percentage of trip distance on different transport modes (Fig. 18)**

Trip type	Mode of transportation					
	Walking		Car		Train	
	Mean	Std.	Mean	Std.	Mean	Std.
Short	72.46	40.61	20.42	35.90	7.12	22.67
Medium	12.67	18.38	79.50	25.15	7.83	17.76
Long	3.27	4.02	95.33	5.79	1.40	3.47
Overall	21.52	32.32	71.51	34.89	6.97	17.82



(a) Overall distribution of distance portion traveled by different modes of transportation

(b) Portion of trip distance traveled by different modes of transportation in a short trip

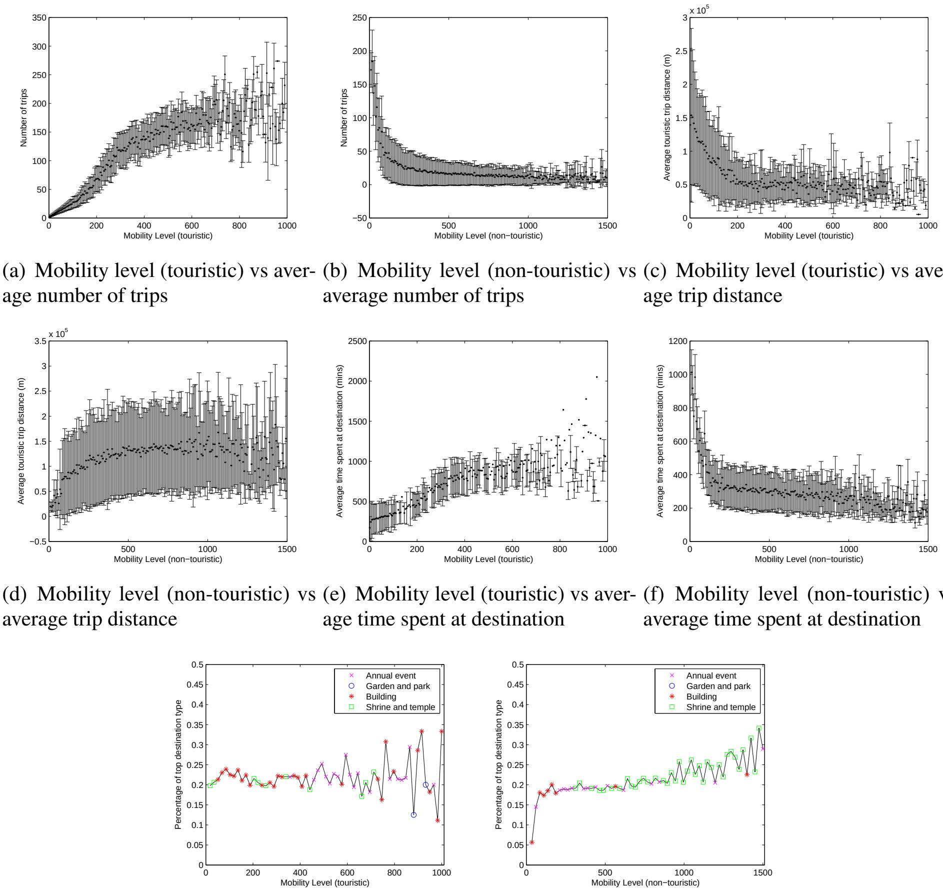
(c) Portion of trip distance traveled by different modes of transportation in a medium trip

(d) Portion of trip distance traveled by different modes of transportation in a long trip

Figure 18: **Distributions of portion of trip distance traveled by train, car, and walking in a trip.**

Table 6: **R-squared values and linear regression constant α of the results shown in Fig. 19, 20, and 21 (ML = mobility level, TD = travel dispersion, TS = travel scope)**

Personal mobility	Tourist travel behavior proxies					
	Touristic			Non-touristic		
	Number of trips	Trip distance	Time spent at destination	Number of trips	Trip distance	Time spent at destination
ML	r^2=0.902 α=0.247	r^2=-0.629 α=-71.683 For ML>200, r^2=-0.238 α=-67.467 For ML≤200, r^2=-0.953 α=-574.33	r^2=0.852 α=1.418	r^2=-0.562 α=-0.013 For ML>200, r^2=-0.745 α=-0.012 For ML≤200, r^2=-0.878 α=-0.348	r^2=0.443 α=126.93	r^2=-0.657 α=0.255 For ML>200, r^2=-0.735 α=-0.249 For ML≤200, r^2=-0.883 α=-3.126
TD	r^2=-0.862 α=-0.077	r^2=0.632 α=-71.683 For TD>80, r^2=0.444 α=6105.7 For TD≤80, r^2=0.957 α=6457.6	r^2=0.545 α=5.58	r^2=0.582 α=0.582	r^2=-0.544 α=2085.1	r^2=0.448 α=9.018
TS	r^2=-0.0.048 α=0.00001	r^2=0.581 α=0.192	r^2=0.505 α=0.00028	r^2=-0.290 α=0.0001	r^2=0.866 α=0.391	r^2=0.074 α=0.0014



(a) Mobility level (touristic) vs average number of trips

(b) Mobility level (non-touristic) vs average number of trips

(c) Mobility level (touristic) vs average trip distance

(d) Mobility level (non-touristic) vs average trip distance

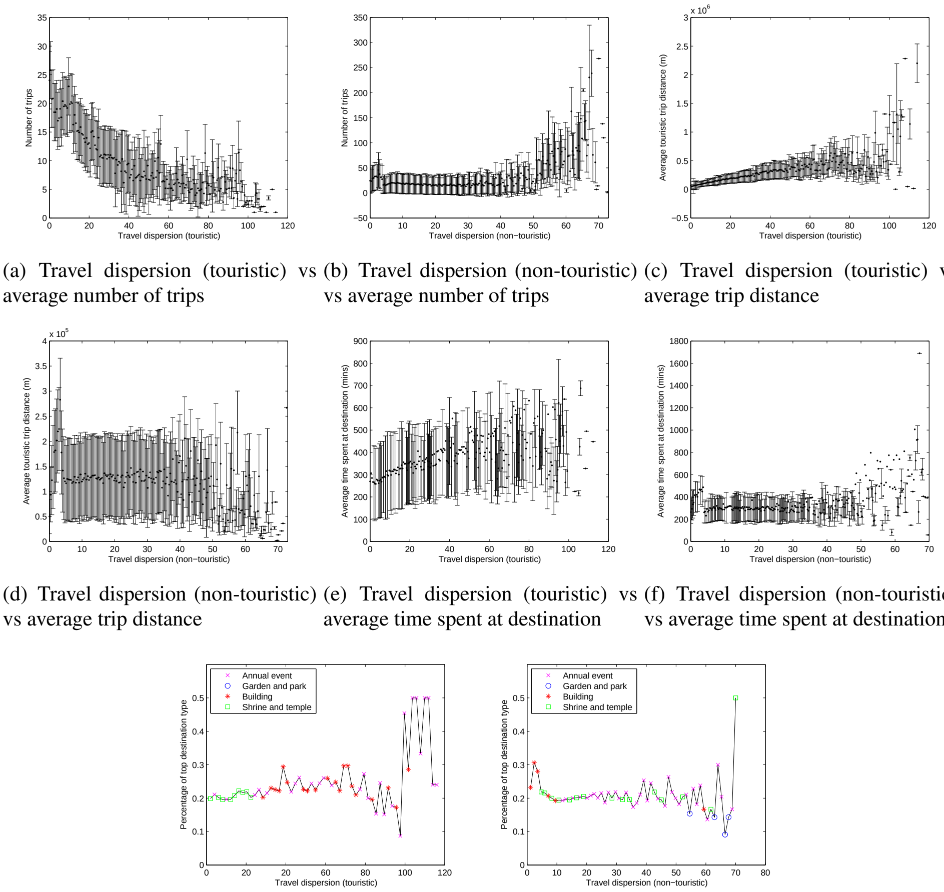
(e) Mobility level (touristic) vs average time spent at destination

(f) Mobility level (non-touristic) vs average time spent at destination

(g) Mobility level (touristic) vs top destination types

(h) Mobility level (non-touristic) vs top destination types

Figure 19: **Mobility level and touristic behavior.**



(a) Travel dispersion (touristic) vs average number of trips

(b) Travel dispersion (non-touristic) vs average number of trips

(c) Travel dispersion (touristic) vs average trip distance

(d) Travel dispersion (non-touristic) vs average trip distance

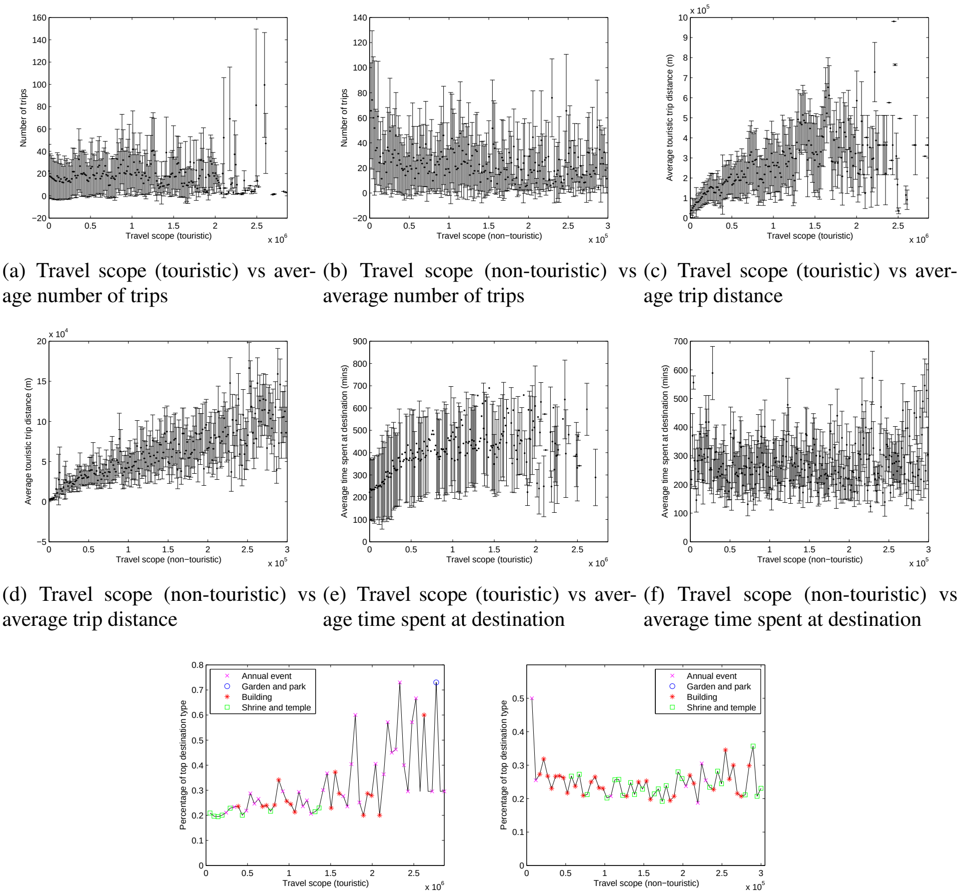
(e) Travel dispersion (touristic) vs average time spent at destination

(f) Travel dispersion (non-touristic) vs average time spent at destination

(g) Travel dispersion (touristic) vs top destination types

(h) Travel dispersion (non-touristic) vs top destination types

Figure 20: **Travel dispersion and touristic behavior.**



(a) Travel scope (touristic) vs average number of trips

(b) Travel scope (non-touristic) vs average number of trips

(c) Travel scope (touristic) vs average trip distance

(d) Travel scope (non-touristic) vs average trip distance

(e) Travel scope (touristic) vs average time spent at destination

(f) Travel scope (non-touristic) vs average time spent at destination

(g) Travel scope (touristic) vs top destination types

(h) Travel scope (non-touristic) vs top destination types

Figure 21: **Travel scope and touristic behavior.**

4. Similarity in travel behavior

Looking into a smaller scale analysis of travel behavior, here we examined the cluster of people based on their travel behavior. It is interesting and also important for urban planning to gain an understanding of people's travel behavior and their clusters within a prefecture i.e., to have an insight into where people share common behavior and what the behavior is.

We considered the *trip distance*, *total time spent at destinations*, *time spent on walking*, *distance traveled by walking*, *time spent traveling by a car*, *distance traveled by a car*, *time spent traveling on a train*, and *distance traveled by a train* as the features to represent the characteristics of travel behavior for each of the 4,734878 total trips made. For each prefecture, we clustered the trips based on these features using the k-means clustering algorithm with $k = 4$. (The number of clusters k was chosen arbitrarily for this preliminary observation of clusters. Other clustering algorithms that do not require pre-assignment of the number of clusters such as DBSCAN is not suitable here due to the high number of data instances (4,734878)).

Once all trips had been clustered, each subject was then assigned to a cluster according to the majority vote scheme based on the subject's clustered trips. If there was a tie, the subject was assigned randomly to one of the clusters of the subject's clustered trips.

From the clustering result that we obtained, we found that for most prefectures, a high number of subjects (more than 70%) in a prefecture were clustered together, which implies that most people have similar travel behavior within the prefecture. Nonetheless there were some prefectures that people were less skewly clustered. Figure 22 shows the portion of people in the prefecture who were assigned to the top cluster (i.e., the most populated cluster).

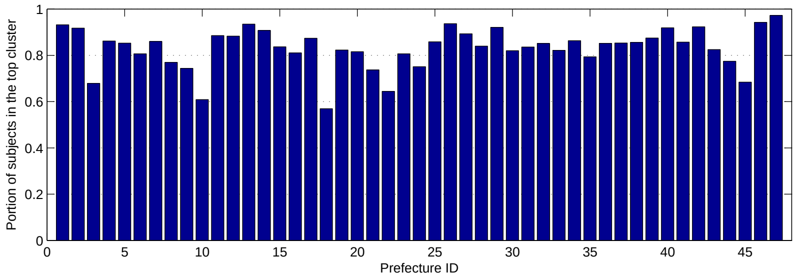
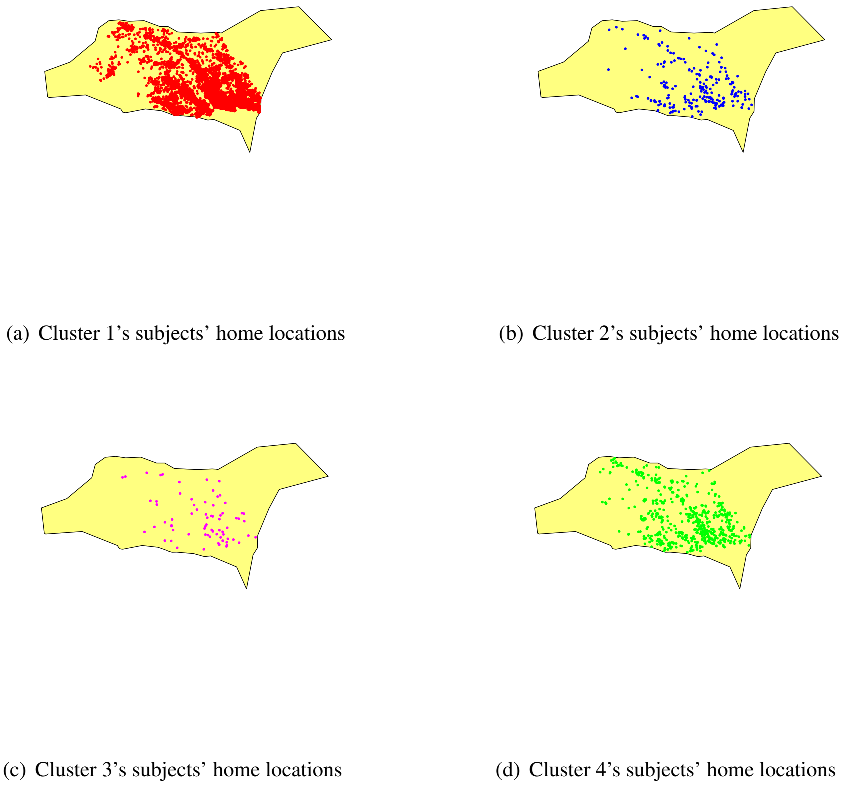


Figure 22: **Portion of the number of subjects who were assigned to the top cluster for each prefecture.**

Geographically, the clustering results did not clearly show that people's travel behavior was strongly influenced by the area of residence. As for examples, subjects' home locations of each cluster for Saitama (prefecture ID 11, where nearly 90% of people's travel behavior are similar) and Fukui (prefecture ID18, where less than 60% of people's travel behavior are similar) prefectures are shown in Fig. 23 and Fig. 24, respectively. People who have similar travel behavior appear to reside in different areas across the prefecture. Yet, for future studies, it will be interesting to further explore the influence of the residential area on people's travel behavior, for example, examining detailed characteristics of area with respect to transport accessibility or urban infrastructure.

Another interesting aspect of the similarity in travel behavior is the characteristics of trips made from one prefecture to another and even within the same prefecture i.e., the O-D matrix



(a) Cluster 1's subjects' home locations

(b) Cluster 2's subjects' home locations

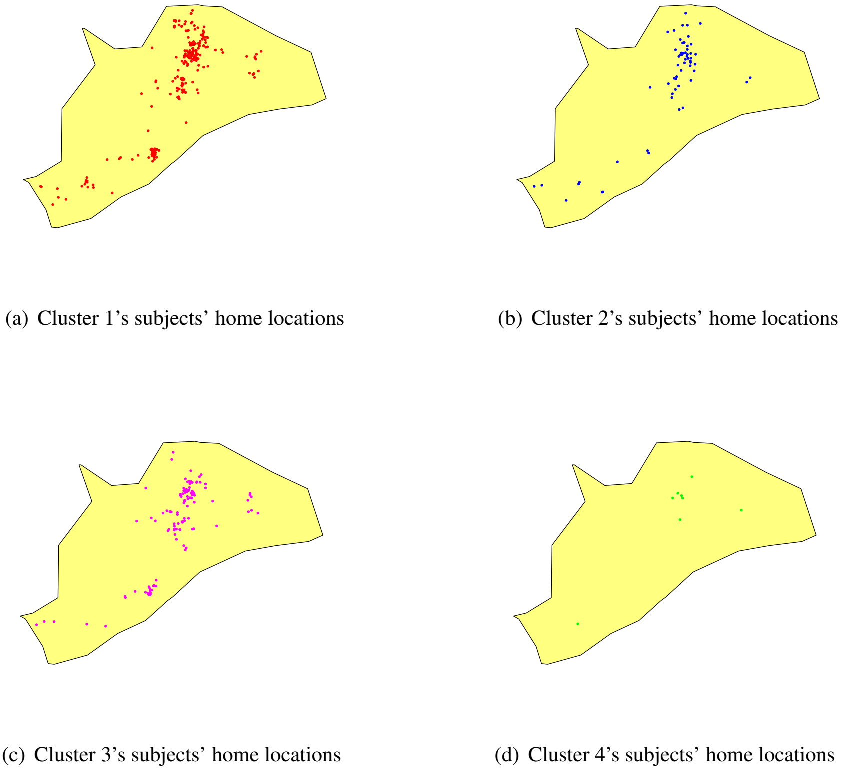
(c) Cluster 3's subjects' home locations

(d) Cluster 4's subjects' home locations

Figure 23: **Home locations of subjects in different clusters in Saitama.**

(Fig. 11). Using the same set of features above we measured the similarity of trips in the O-D matrix by firstly normalizing each feature (separately) to [0,1], then calculating the standard deviation for each normalized feature i (S^i_{nor}) from which the *trip similarity* can be calculated as $1 - mean(\{S^1_{nor}, S^2_{nor}, ..., S^8_{nor}\})$.

Figure 25 shows the O-D similarity matrix from which we observed that the similarity of trips within the same prefecture is generally higher than trips made to and from a different prefecture. It can also be observed that the O-D similarity matrix is similar to the O-D matrix (shown in Fig. 11). The *correlation coefficient* (a measure of the linear correlation) between the O-D matrix and the O-D similarity matrix was calculated to be 0.6769, suggesting that there is some correlation between the number of trips and trip similarity. In other words, there is a tendency that when the destination becomes more popular (i.e., attracting a higher number of visits (or trips)), visitors (both new and repeat) tend to travel to the destination in a similar way – presumably following suggestions by other people who have previously visited the destination (for the new visitors) or repeating the same traveling pattern to the destination (for the repeat visitors). When the destination becomes popular, this tends to be the case that there are a few popular ways to get to the destination.



(a) Cluster 1's subjects' home locations

(b) Cluster 2's subjects' home locations

(c) Cluster 3's subjects' home locations

(d) Cluster 4's subjects' home locations

Figure 24: **Home locations of subjects in different clusters in Fukui.**

5. Application

Based on our framework, a number of tourism applications can be developed. Here, we demonstrate an application that can be useful, particularly for urban planners, transport management, and tourism authorities. We developed a user interface that allows the user to observe and analyze mobility patterns of people, in particular, travel behavior. We would like to demonstrate it at two levels: individual and aggregate. At the individual level, a user can observe touristic, non-touristic, and commuting trajectories, all of which allow the user to conduct further investigations into travel behavior, or to make more informed decisions in urban planning and transportation. A snapshot of this user interface is shown in Fig. 26. As an example, the trajectories of four subjects in the Tokyo area are shown, along with the types of trips and the locations of home and workplace.

At the aggregate level, a user can observe and analyze tourist flows across different cities or prefectures. The user interface allows the user to observe the inflow and outflow of any prefectures of interest and for any selected period of observation in order to analyze the trends and flow distributions. A snapshot our user interface is shown in Fig. 27. In this example, the outflow of Tokyo is illustrated.

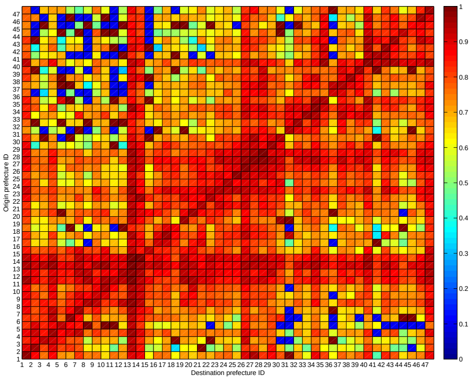
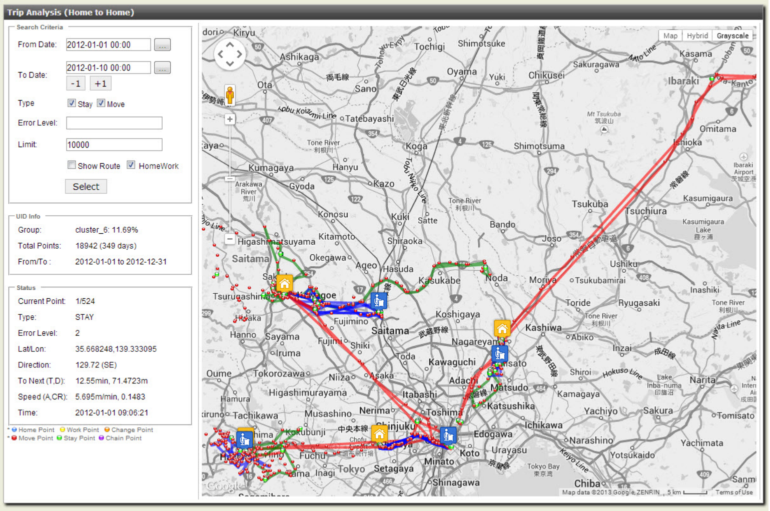


Figure 25: **O-D similarity matrix.**

6. Conclusion

The ubiquity and sensing capabilities of personal mobile phones can transform such phones into human probes – individual sensors that can monitor and record individual behavioral data, which collectively can help reveal interesting human behavior phenomena that govern how cities function and how different urban elements operate. We recognized and capitalized on this opportunistic sensing mechanism and carried out a study of tourist behavior using considerable mobile phone GPS location records collected from 130,861 users in Japan for a full calendar year. We described our algorithm to detect tourists from GPS traces that allowed us to conduct our study. We particularly investigated the number of trips made, the time spent at destinations, the modes of transportation used, and the relationship between personal mobility and tourist travel behavior. The main findings of our study can be summarized as follows:

- The top points of origin for tourists Tokyo, Kanagawa, and Fukuoka and most trips were medium length trips (5-100km). The top destinations were Tokyo, Kyoto, and Fukuoka; most trips were a mixture of short (less than 5km) and medium length trips.
- Large tourist flows were observed between Tokyo and Kanagawa, as well as between Fukuoka and Yamaguchi. High trip volume was observed between prefectures that are close to each other and within the same region. Prefecture population density appeared to be indicative of outflow volume, but not inflow volume.
- Tourists were more likely to spend more time (nine hours or more) at destinations on longer trips.
- Shrine and Temple, Building, and Annual Events were the top types of destination that drew a large number of tourists. Open Field, Botanical Garden and Aquarium, and Museum were the top destination types that tourists spent most of their time on a trip.
- As trips became longer, more time was spent in a car, but less time was spent walking. Time spent on a train was the least among other modes of transportation.
- Overall, cars were used to travel more distances on a trip than trains and walking. As trips became longer, cars were used to travel longer distances, whereas distances traveled by



Trip Analysis (Home to Home)

Search Criteria

From Date: 2012-01-01 00:00

To Date: 2012-01-10 00:00 -1 +1

Type ☑ Stay ☑ Move

Error Level:

Limit: 10000

□ Show Route ☑ HomeWork

Select

UID Info

Group: cluster_6: 11.69%

Total Points: 18942 (349 days)

From/To : 2012-01-01 to 2012-12-31

Status

Current Point: 1/524

Type: STAY

Error Level: 2

Lat/Lon: 35.668248,139.333095

Direction: 129.72 (SE)

To Next (T,D): 12.55min, 71.4723m

Speed (A,CR): 5.695m/min, 0.1483

Time: 2012-01-01 09:06:21

Figure 26: **A snapshot of our user interface that allows the user to observe different types of trips. This example shows four subjects' trajectories along with the type of trips made and locations of home and workplace. Commuting trips are represented in blue, touristic trips are represented in red, and non-touristic trips are represented in green.**

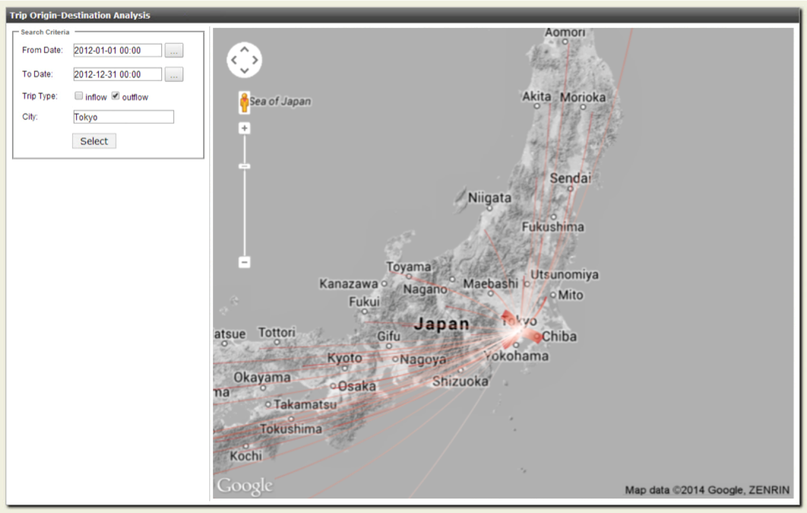
walking decreased. Train remained the least used mode of transportation among others in terms of distance traveled.

- Tourists who were more frequent travelers were more likely to spend more time at destinations.

- Tourists whose travels were more dispersed were more likely to make smaller number of trips and vice versa.

- People who visited places locater farther for non-touristic purposes were more likely to make longer touristic trips.

In addition, we developed an application that allows users to observe people's mobility that can be classified into commuting, non-touristic, touristic trips, as well as the locations of people's home and workplace. Our application also allows users to observe and analyze inflow and outflow of any prefectures of interest within any selected period of observation. We believe that the application can be useful for urban planners, transportation management, and tourism authorities. Nonetheless, there are a number of limitations of this study. First, the inferred touristic trips might not have included all actual touristic trips. Second, other modes of transportation, such as cycling and bus, were not considered in the analysis. Last, travel scope was considered only for domestic trips; hence, international trips were excluded from the analysis. This study offers another way of understanding tourist behavior with an advantage over traditional surveys and questionnaire studies in terms of the size and longitudinality of the data for the analysis. The findings we present also offer interesting insights into tourist travel behavior. Our future direction for this research trajectory includes real-time sensing and ICT-driven mechanisms for sustainable tourism.



Trip Origin-Destination Analysis

Search Criteria

From Date: 2012-01-01 00:00

To Date: 2012-12-31 00:00

Trip Type: ☐ inflow ☑ outflow

City: Tokyo

Figure 27: **A snapshot of our user interface that allows the user to observe tourist flows. This example shows outflow of Tokyo.**

Acknowledgments

This study has been supported by the JSPS (Japan Society for the Promotion of Science) Invitation Fellowship (S-13176), the DIAS (Data Integration and Analysis System) and GRENE (Green Net- work of Excellence) project of MEXT (Ministry of Education, Culture, Sports, Science and Technology). Special thanks to ZENRIN DataCom CO., LTD for the dataset used in this study.

References

[1] S. Phithakkitnukoon, T. Horanont, G. D. Lorenzo, R. Shibasaki, C. Ratti, Activity-aware map: Identifying human daily activity pattern using mobile phone data, in: Human Behavior Understanding, 2010, pp. 14–25.

[2] M. C. González, C. A. Hidalgo, A.-L. Barabási, Understanding individual human mobility patterns, Nature 458 (2009) 238–238. doi:10.1038/nature07850.

[3] C. Song, Z. Qu, N. Blumm, A.-L. Barabási, Limits of predictability in human mobility, Science 327 (5968) (2010) 1018–1021.

[4] C. Song, T. Koren, P. Wang, A.-L. Barabási, c, Nature Physics 6 (10) (2010) 818–823.

[5] S. Phithakkitnukoon, Z. Smoreda, P. Olivier, Socio-geography of Human Mobility: A study using longitudinal mobile phone data, PLoS ONE 7 (6).

[6] J. Alegre, L. Pou, The length of stay in the demand for tourism, Tourism Management 27 (6) (2006) 1343–1355.

[7] U. Gokovali, O. Bahar, M. Kozak, Determinants of length of stay: A practical use of survival analysis, Tourism Management 28 (3) (2007) 736–746.

[8] L. Wu, J. Zhang, A. Fujiwara, Dynamic analysis of japanese tourists' three stage choices tourism participation, destination choice, and travel mode choice, Transportation Research Record 2322 (2012) 91–101.

[9] Y. Zheng, Q. Li, Y. Chen, X. Xie, W.-Y. Ma, Understanding mobility based on gps data, in: Proceedings of the 10th International Conference on Ubiquitous Computing, UbiComp '08, ACM, New York, NY, USA, 2008, pp. 312–321.

[10] M. Ester, H. peter Kriegel, J. S, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), AAAI Press, 1996, pp. 226–231.

[11] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32.

[12] Statistics Bureau, Ministry of Internal and Communications, The 2006 Establishment and Enterprise Census, [Online; accessed 10-July-2013].
URL `http://www.stat.go.jp/english/data/jigyou/2006/index.htm`

[13] Japanese Ministry of Land, Tourism Resources, [Online; accessed 15-Nov-2013].
URL `http://nlftp.mlit.go.jp/ksj/gml/gml_datalist.html`

[14] L. Breiman, Degrowing tourism: Decroissance, sustainable consumption and steady-state tourism, Anatolia: An International Journal of Tourism and Hospitality Research 20 (1) (2009) 46–61.

[15] A. Witayangkurn, T. Horanont, N. Ono, Y. Sekimoto, R. Shibasaki, Trip reconstruction and transportation mode extraction on low data rate gps data from mobile phone, in: Proceedings of the International Conference on Computers in Urban Planning and Urban Management (CUPUM 2013), 2013, pp. 1–19.